Amendments to the Specification:

Please replace paragraph [0005] with the following amended paragraph:

[0005] Some embodiments of the computer implemented methods of the invention include the steps of obtaining a plurality of models (such as a battery of HMMs), where each of the models represents a classification of biological sequences with structural or functional similarity; determining distances of the biological sequences to the models; and automatically classifying the sequences according to the distances to the models. The methods are particularly suitable for annotating a large number of sequences, preferably at least 50, 100, 500, 1000, or 5000 sequences. In preferred embodiments, the sequences are protein sequences. In such embodiments, the models are typically established according to structural relationships among known proteins. Data from many protein databases, such as the Structural Classification of Proteins (SCOP) (~~http://~~ World Wide Web Address: scop.berkeley.edu) and Protein Data Bank (PDB) (~~http://~~ World Wide Web Address: www.rcsb.org/pdb/) are useful for building the HMM libraries in at least some embodiments of the invention.

Please replace paragraphs [0026] and [0027] with the following amended paragraphs:

[0026] FIGURE 3 shows an exemplary computer network that is suitable for executing the computer software of the invention. A computer workstation 302 is connected with the application/data server(s) through a local area network (LAN) 301, such as an Ethernet 305. A printer 304 may be connected directly to the workstation or to the Ethernet 305. The LAN may be connected to a wide area network (WAN), such as the Internet 308, via a gateway server 307 which may also serve as a firewall between the WAN 308 and the LAN 305. In preferred embodiments, the workstation may communicate with outside data sources, such as the National Biotechnology Information Center, through the Internet. Various protocols, such as FTP and HTTP, may be used for data communication between the workstation and the outside data sources. Outside genetic data sources, such as the GenBank 310, are well known to those skilled in the art. An overview of GenBank and the National Center for Biotechnology information

(NCBI) can be found in the web site of NCBI (~~http://~~ World Wide Web Address: www.ncbi.nlm.nih.gov). Other useful Internet accessible databases include Structural Classification of Proteins (SCOP) (~~http://~~ World Wide Web Address: scop.berkeley.edu) and Protein Data Bank (PDB)(~~http://~~ World Wide Web Address: www.rcsb.org/pdb/). Additional biological databases accessible through the internet are described in, *e.g.*, Special Issue on Biological Databases, Nucleic Acid Research, 2000, 29:0-349, incorporated herein in its entirety by reference for all purposes. Protein databases are also discussed in Section II, *infra*.

[0027] Computer software products of the invention typically include computer readable medium having computer-executable instructions for performing the logic steps of the methods of the invention. Suitable computer readable medium include floppy disk, CD-ROM/DVD/DVD-ROM, hard-disk drive, flash memory, ROM/RAM, magnetic tapes and etc. The computer executable instructions may be written in any suitable computer language or combination of several languages. Suitable computer languages include C##, C/C++ (such as Visual C/C++), Java, Basic (such as Microsoft Visual Basic), SQL, Fortran, SAS and Perl. The software may be written in the form of independent software, software components (such as Java beans, Enterprise Java Beans). Software of the invention may also be implemented to provide application programming interfaces (APIs). In preferred embodiments, the software of the invention may provide remote services using remote execution in a distribute fashion. For example, in some embodiments, some or all of the logic steps of the methods and software of the invention are implemented as Web Services. For a detailed description of Web Services, see, e.g., "Web Services Insider," by James Snell, available at: ~~http://~~ World Wide Web Address: www-106.ibm.com/developerworks/webservices/library/we-ref1.html, last visited September 20, 2001.

Please replace paragraph [0053] with the following amended paragraph:

[0053] *Whole genome gene set.* A set of protein sequences covering the Golden Path of the human genome (October 7, 2000 freeze ~~http://~~ World Wide Web Address: genome.ucsc.edu/)

was generated by the Genie (Reese, M. G., Kulp, D., Tammana, H. and Haussler, D. (2000) Genie--gene finding in Drosophila melanogaster. Genome Res 10(4), 529-38) programs suite (Kulp, D. & Wheeler, R., available at the following URL (genome.ucsc.edu, last visited on September 20, 2001), with the repeat regions masked out (802). The data set consists of three sets of amino acid sequences: (1) the set derived by alignment of mRNA to the genomic DNA corresponding to sequences in RefSeq and GenPept, (2) a set of alternatively spliced variants, which are generated using mRNA/EST-to-genomic alignments in combination with purely statistical methods, each of which contain the largest subset of exons for a particular gene, and (3) genes predicted by purely ab initio methods. These sequences are non-redundant; none of the included genes overlap the same genomic region. In cases where there were many genes overlapping the same region, the one with the longest CDS (translation) was kept. This set, known as annot10, contains 59,378 putative protein sequences.